

基于权重的 Apriori 算法在文本统计特征提取方法中的应用*

李昌兵 庞崇鹏 李美平
(重庆邮电大学经济管理学院 重庆 400065)

摘要: 【**目的**】解决在海量客户评论信息中抽取产品特征时噪声大的问题。【**方法**】运用 TF-IDF 和方差选择的统计方法在众多初步提取出来的特征中进行选择, 设置阈值后将各自提取出来的特征取交进行过滤, 得到产品特征集合, 根据基于矩阵和权重改进的 Apriori 算法产生频繁项集, 设定不同阈值得到最优特征集合, 实现对用户评论中产品特征的自动提取。【**结果**】以手机评论文本为例, 从中抽取手机类的产品特征, 根据人工标注的 183 个特征和算法识别出来的特征, 查准率 P 为 72.44%, 查全率 R 为 77.59%, 综合值 F 为 74.93%。【**局限**】查准率偏低, 存在人工标注特征错误的情况。【**结论**】实验结果表明, 在用统计方法和改进后的 Apriori 算法进行特征提取时可以提高各性能指标。

关键词: 特征提取 Apriori 算法 TF-IDF 方差选择
分类号: G350

1 引 言

随着互联网的普及, 网络产品评论数量飞速增长, 很多企业已经在逐渐将重心转移到数据领域。通过人工处理方式从这些产品评论文本信息中获取有用的信息越来越困难。因此, 借助一定技术手段实现这一过程变得尤为重要。

产品特征包括产品属性以及构成产品的各个方面, 可使用户方便快速地了解到产品的特点。如功能、屏幕、图片、价格等手机类产品特征。现如今, 许多国内外学者在特征挖掘的研究中已经取得了一些成果。Zhuang 等[1]采用人工或半自动的方式对电影中文评论领域进行产品特征提取研究。Kobayashi 等[2]提出利用产品、产品特征和观点词之间的共现模式的半自动化方法提取产品特征和观点词。娄德成等[3]利用半自动方式进行人工定义, 从而抽取出产品评论信息。Hu 等[4]抽取出现频率大的名词及名词短语作为候选产品特征, 通过压缩剪枝和冗余剪枝策略对提取的频繁商品特征进行筛选, 再使用关联规则挖掘识别频繁产品特征。此方法使得各性能指标有了较大提升。Popescu 等[5]将产品特征看作是产品的一部分, 使用候选产品特征和领域特征之间的共现提取商品特征, 并使用点互信息 PMI(Pointwise Mutual Information)表示关联程度, 最终按关联程度大小选择商品特征。该方法提高了产品特征提取的准确率, 但召回率有所下降。随着关联规则算法 Apriori 与 FP 在数据挖掘和机器学习领域不断被应用, 旨在挖掘出事物项之间的内在联系, 这两种算法也被应用于特征频繁项集挖掘,

通讯作者: 庞崇鹏, ORCID: 0000-0002-4559-0556, E-mail: Pang_Aaron@163.com。

*本文系国家自然基金项目"基于群体智能的多 Agent 协作模型与适应性研究"(项目编号: 60905066)、重庆邮电大学自然科学基金资助项目"时间序列数据挖掘技术应用研究"(项目编号: A2009-03)和电子商务与现代物流重庆市高校市级重点实验室重点项目"基于多主体博弈的供应链契约选择与协调控制机制研究"(项目编号: ECML201403)的研究成果之一。

并且取得了理想的效果。然而采用传统的 Apriori 算法进行特征提取也存在一些不足：杜思奇等[6]先利用 Apriori 算法产生频繁集再用 TF-IDF 阈值进行过滤，准确率得到了较大提升，但是使用 Apriori 算法初步产生频繁项集会带来许多的非产品信息，特别是在评论语料大的情况下，导致性能指标有所下降。王永等[7]利用 FP 增长算法产生频繁项集，根据独立支持度、频繁项名词非特征规则及 PMI 阈值过滤技术对候选产品特征进行筛选。文中在用 FP 算法时采用最小支持度 1%进行实验，支持度设置的越小，查全率也就会越高，但是在产生的频繁项集中噪声也就相应越大，在后续的工作中也会带来较大干扰。路永和等[8]综合分析特征提取方法并对传统特征提取流程和方法进行改进，利用特征池进行特征词预选，再引入遗传算法对候选特征词分组编码并提取最佳特征向量。在特征预选阶段采用特征选择方法 CHI 和 IG，通过比较去重形成特征池。但是这样会造成一个问题就是重复的特征大部分是重要的特征，两种方法提取出的结果中未重复的特征大部分为非重要特征，再取则会将那些不重要的特征集合进一步扩大。这些方法虽然在一定程度上使得特征提取方法的各性能指标有所提升，但是在评论语料足够多的情况下不利于噪声的清除，中文产品评论领域特征提取的挖掘性能也有待进一步提高。

鉴于此，本文在特征预选阶段采用了特征选择方法方差分析与 TF-IDF 方法进行取交操作形成候选特征集合，然后采用基于矩阵和权重的改进 Apriori 算法进行频繁项集挖掘，此改进算法可避免数据库的重复扫描，使得时间和空间的耗费显著减少，同时能有效的挖掘出更有价值的事件。为了验证该方法的有效性，本文以手机类产品评论为例进行特征抽取。

2 产品特征提取流程

在现有的许多产品特征提取方法上，产品特征一般提取流程如图 1 所示。与前人研究相比较，本文为降低噪声数据的比例，在对特征预抽取方法上进行改进，在特征预抽取阶段采用基于方差分析与 TF-IDF[9]方法进行特征预选择，分别筛选出排名前 1 000 的特征集合进行交操作。制定名词非特征规则，建立相应名词集合进一步筛选产品特征；利用基于矩阵和权重的改进 Apriori 算法[10]，设定最优阈值，形成最终的产品特征集合。

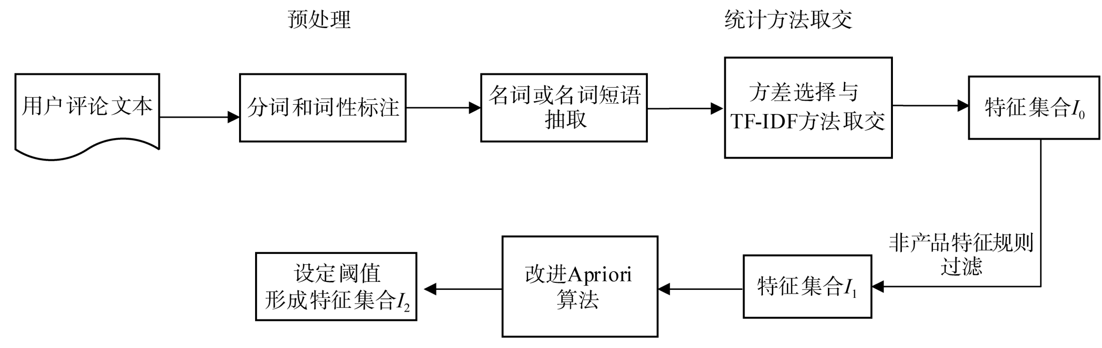


图 1 基于统计和权重的产品特征提取流程

(1) 应用 Python 工具的 jieba 分词包对原始评论语料进行分词和词性标注。

(2) 根据 jieba 分词工具所使用的词语标记符号，其中与名词相关的子集标记符号有{/n, /nr, /ns, /nt, /nz, /nl, /ng}，再根据这些标记符号所代表的含义和语法特点，本文选取{/n}作为抽取规则。使用计算机程序对每一条评论进行抽取。

(3) 采用方差选择法和 TF-IDF 对初步抽取出来的特征进行预选择，再分别选取排名前 1 000 的特征；将两种方法抽取出的特征取交集得到产品特征集合 I_0。

(4) 建立常见中文频繁项名词却非产品特征的集合，并从中文语义及语法角度过滤 I_0，形成特征集合 I_1。

(5) 常见的频繁项名词却非产品特征主要划定为以下情况。

①常见商品的品牌。例如“诺基亚”、“三星”、“西门子”等名词。

②一些常见的口语化名词。例如“机子”、“情况”、“卖点”、“优缺点”等。

③与产品无关的称呼类名词，例如“朋友”、“同事”、“男

子”等。

④计算机程序识别出来的少量错误名词，例如“高端”、“聊天”、“海量”等。

⑤常见的集合类名词，例如“群组”、“大家”等。

(6) 采用基于矩阵和权重的改进 Apriori 算法设置最优阈值提取最终特征集合。

3　方法设计

在本文方法中，为了在特征预抽取阶段避免特征维度过高而导致噪声数据带来的影响，而选取方差分析与 TF-IDF 这两种方法进行候选特征提取。方差分析适用于特征值都为离散型的变量，符合本文构建的数据结构 DataDframe。同时对于用于机器学习的数据来说，方差大才有意义，包含的信息量也就越大，并且通过实验结果也可以看出，方差越大的特征提取效果越优。而 TF-IDF 算法则是通过加权判定特征项对于评论语料的重要性，旨在过滤常用词。比如在“手机”，“国产”和“功能”出现频次相同情况下，明显“功能”更为重要。并且在前人研究中此算法在特征提取领域中也表现出了较好的挖掘性能。同时，针对 TF-IDF 算法对文章不同位置的词语一视同仁这个不足之处，在本文中，评论文本多为短文本，所以将 TF-IDF 用于短文本特征挖掘也是行之有效的。实验结果显示该方法特征提取效果也较为明显。

在进行以上两种方法取交过滤后，本文同时也引入一种在特征提取领域研究者尚未采用的基于矩阵与权重的改进 Apriori 算法。此改进算法主要是基于事物项的权重而提出的，跟传统 Apriori 算法相比，避免了数据库的重复扫描，并且能够有效挖掘出潜在且更有价值的事件。

3.1　结合方差分析与 TF-IDF 算法

本文对特征选择方法 PMI、TF-IDF、TF-IWF 和方差分析 4 种方法进行实验对比分析，选取其中效果较好的 TF-IDF 与方差分析两种方法进行本文产品特征预抽取。

(1) 方差选择法：将评论语料和特征转成字典形式，利用 key 值构建数据结构 DataDframe，评论语料为行索引值，特征为列索引值。行索引集合为{T_1, T_2, T_3, …, T_m}，列索引值为{A_1, A_2, A_3, …, A_n}。其中 m 为产品评论语料数量，n 为特征数量。用 0-1 填充 DataDframe，1 代表特征 A_n 在相应评论语料 T_m 里面，0 代表特征 A_n 不在相应评论语料 T_m 里面，再对每一列的特征求方差。数据结构 DataDframe 形式如表 1 所示。

表 1　数据结构 DataDframe

	A_1	A_2	A_3	…	A_n
T_1	0	0	1	…	1
T_2	0	1	0	…	0
T_3	1	1	0	…	1
…	…	…	…	…	…
T_m	1	1	1	0	0

(2) TF-IDF 选择法：用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。在一份给定的文件里，词频 (Term Frequency, TF)指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文件。逆向文件频率 (Inverse Document Frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的结果取对数得到。某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。TF-IDF 的计算如公式(1)-公式(3)所示。

$$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

其中，$n_{i,j}$ 是该词在评论语料 d_j 中的出现次数；而分母则是在评论语料中所有字词的出现次数之和。

$$IDF_i=\log\frac{|D|}{\left|\{j:t_i\in d_j\}\right|} \qquad (2)$$

其中，$|D|$ 是语料库中的评论总条数；$\left|\{j:t_i\in d_j\}\right|$ 是包含词语的文件数目，如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用 $1+\left|\{j:t_i\in d_j\}\right|$。

$$TF-IDF=TF_{i,j}\times IDF_i \qquad (3)$$

基于以上两种方法，本文在特征预抽取阶段结合方差分析与 TF-IDF 分别进行特征抽取，然后取出维度为 1 000 的特征进行取交操作形成产品特征集 I_0。

3.2　基于矩阵与权重的改进 Apriori 算法

本文将基于矩阵与权重的改进 Apriori 算法应用到文本挖掘领域，通过实验结果分析，该算法使得本

文特征抽取效果得到了较大提升。算法设计如下:

用评论语料和特征集合 I_l 构建 0-1 矩阵 M:

$$M=\begin{Bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{Bmatrix}$$

其中，$a_{ij}=\begin{cases}1, a_{ij}\in T_i \\ 0, a_{ij}\notin T_i\end{cases}$；$T_i$ 表示第 i 条评论; i=1, 2, 3⋯, m; j=1, 2, 3, ⋯, n; I={I_1，I_2，I_3，⋯, I_N} 表示 N 个特征集合。I_j 在事务数据库中出现的概率为 $p(I_j)$，计算如公式(4)所示，I_j 的权重记为 $w(I_j)$，与 $p(I_j)$ 有关，$w(I_j)$ 的计算如公式(4)–公式(5)所示。

$$p(I_j)=l/m \tag{4}$$

$$w(I_j)=1/p(I_j) \tag{5}$$

其中，l 表示 I_j 在事务集中出现的次数，即上述矩阵中第 j 列中 1 的个数, m 是评论语料的总条数。

事务 T_k 指数据集中的第 k 条评论，其权重指该评论中所包含的特征项的平均权重，记为 $wt(T_k)$，即对 $a_{ij}=1$ 的所有 $w(I_j)$ 求平均值，其中 $j=1,2,3,\cdots,n$，计算如公式(6)所示。

$$wt(T_k)=\sum_{j=1}^{I_j\in T_k} w(I_j)/|T_k| \tag{6}$$

其中，$|T_k|$ 表示评论 T_k 中包含的特征项的个数。

项的权重支持度记为 *wsupport*，权重支持度表示包含特征项的事务权重占所有事务权重的比例，再根据特征项的权重支持度，设定合理阈值形成最优特征集合，计算如公式(7)所示。

$$wsupport(\mathrm{S})=\sum_{k=1}^{S\subseteq T_k} wt(T_k)/\sum_{k=1}^{m} wt(T_k) \tag{7}$$

其中, S 表示事务数据库中的任意特征项。

基于矩阵和权重的改进 Apriori 算法步骤如下:

①扫描事务数据库，构建 0-1 事务矩阵，并根据事务矩阵计算出每个特征项和事务的权重，即 $w(I_j)$，$wt(T_k)$。

②根据事务矩阵得到候选 1-项集 C_1，计算 C_1 中每个特征项的权重支持度 *wsupport*(S)，找出满足最小支持度的频繁 1-项集 L_1。

基于矩阵和权重的改进 Apriori 算法流程图如图 2 所示。

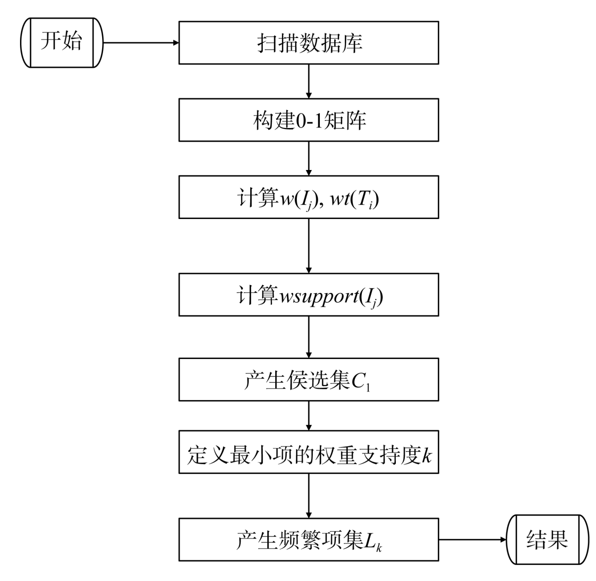


图 2　基于矩阵和权重的改进 Apriori 算法流程

3.3　性能评估指标

本文采用查准率 P、查全率 R 和综合值 *F-score* 这三个评估指标分别来度量性能的某个方面和对性能的整体评估。具体计算方法如公式(8)–公式(10)所示。

$$P=\frac{A}{A+B} \tag{8}$$

$$R=\frac{A}{A+C} \tag{9}$$

$$F\text{-}score=\frac{2RP}{R+P} \tag{10}$$

其中, A, B, C 含义如表 2 所示。

表 2　各变量含义

产品特征数	算法识别出来的正确特征数	算法识别出来的错误特征数
挖掘出的特征数	A	B
没有挖掘出的特征数	C	—

4　实验结果及性能评估

4.1　实验数据

本文数据集采用数据堂提供的手机评论语料[①]，选取其中 800 条评论进行实验。通过人工标注的方法共得到上述评论语料中的手机产品特征 183 个，产品特征集合如表 3 所示。

①http: //www.datatang.com/data/43824.

表 3　手机产品特征

产品名称	参数	人工标注特征集合	人工标注特征数量
手机	外观设计	外键，外屏，彩屏，机身，磨砂，键盘，外观，内屏，方向键，外观设计，颜色，手感，外壳，体积，重量，快捷键，金属，质感，机型，外形，面积，按键，数字键，导航键，造型，功能键，机体，材质，图案，拨号键，外表，数字键盘，红外接口，尺寸，按钮，外盖，机壳	37
	屏幕	分辨率，色彩，屏保，画面，屏幕，清晰度，亮度，屏幕显示，显示屏，触摸屏，画质，动画，透明度	13
	基本功能	功能，短信，通话记录，计算器，记事本，程序，联系人，手写，信息，电话，短消息，彩信，闹钟，日程表，手写输入，语音，软件，收音机，防火墙，通话质量，电话簿，录音，电话号码，号码，输入法，语音拨号，键盘输入，通话，闹铃，通讯录，应用程序，时钟，背光灯，录音器，背景灯，手电筒，备忘录，收件箱，SIM 卡	39
	摄像功能	像素，摄像头，彩灯，图片，闪光灯，照片，象素，镜头，图像，照相机，摄像机	11
	娱乐功能	多媒体，影音，媒体播放器，游戏，音频，播放器	6
	数据功能	蓝牙，红外线	2
	手机附件	耳机，手写笔，扩音器，耳塞，内存卡，存储卡，数据线，充电器，防尘盖，传输线	10
	美化	壁纸，界面，背景，菜单，饱和度，主题	6
	性能	信号，响应速度，速度，识别率，待机时间，续航，性能，处理速度，关机，操作速度，网络，待机，反应速度，开机，传输速度，速率，反应时间，智能，输入速度	19
	声音	铃音，铃声，音量，提示音，声音，和弦，和弦铃声，音质，音乐，听筒，扬声器，音效，短信铃声，关机闹钟	14
	硬件配置	容量，内置，空间，储存量，内存，处理器，电池，硬件，外置，存储量，存储容量，均衡器，电池容量，储存，内存容量，电池电量，存储空间，储存卡	18
	性价比	性价比，价格，价位，价钱，价值，零售价	6
	售后反馈	质量，客服	2

4.2　实验结果

(1) 产品特征提取结果

根据公式(7)计算出各特征项的权重支持度，并提取出排在前 10 的手机特征项，如表 4 所示。

表 4　手机产品特征提取结果

排名	属性	*wsupport*
1	功能	0.3337
2	屏幕	0.2628
3	效果	0.2348
4	铃声	0.2324
5	外观	0.2057
6	电话	0.2054
7	短信	0.1887
8	待机	0.1772
9	声音	0.1719
10	电池	0.1685

对 *wsupport* 设置不同的阈值，性能变化如图 3 所示，相应的性能指标值如表 5 所示。

从表 5 可以看出，*wsupport* 阈值为 0.013 时，挖掘结果综合性能最优，即查准率达到 72.44%，查全率达到 77.59%，综合值达到 74.93%。

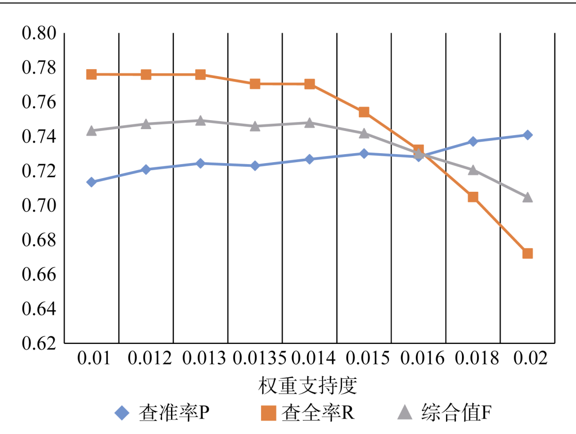


图 3　不同阈值下的性能变化情况

表 5　手机评论挖掘性能

项的权重支持度	P(查准率)	R(查全率)	F(综合值)
0.01	71.35%	77.60%	74.34%
0.012	72.08%	77.59%	74.73%
0.013	72.44%	77.59%	74.93%
0.0135	72.30%	77.05%	74.60%
0.014	72.68%	77.04%	74.80%
0.015	73.01%	75.41%	74.19%
0.016	72.82%	73.22%	73.02%
0.018	73.71%	70.49%	72.06%
0.02	74.09%	67.21%	70.48%

(2) 实验结果对比分析

文献[7]采用 FP 增长算法产生候选特征集，利用基于网络搜索引擎的 PMI 算法进行最优特征提取；文献[11]人工定义了产品属性概念模型，依据此模型对中文产品特征进行提取；文献[13]结合汉语中名词性短语的表达特点，在传统 Apriori 算法基础上进行名词短语扩充，实现产品特征的自动提取。以上三种方法的挖掘性能都有一定提升。将本文方法分别与文献[7]，文献[11]，文献[13]进行比较，结果如表 6 所示。

表 6　针对手机评论的产品特征挖掘结果比较 1

性能指标	本文方法	文献[7]的方法	文献[11]的方法	文献[13]的方法
查准率	72.44%	70.8%	70.72%	62.8%
查全率	77.59%	73.3%	68.35%	81.8%
综合值	74.93%	72%	69.51%	71.05%

通过表 6 可知，本文方法查准率均优于文献[7]、文献[11]和文献[13]；查全率优于文献[7]和文献[11]，但低于文献[13]，由于文献[13]针对的是英文评论，没有绝对的可比性，但是本文挖掘性能更优；从综合性能来看，本文综合评价指标均优于其他文献。通过第一组对比实验可知，利用统计方法和机器学习算法进行产品特征挖掘更有效。

由于文献[12]和文献[4]的方法在中文产品特征提取领域具有一定代表性，因此再将本文方法与其进行实验结果对比，结果如表 7 所示。

表 7　针对手机评论的产品特征挖掘结果比较 2

性能指标	本文方法	文献[12]的方法	文献[4]的方法
查准率	72.44%	63.3%	71.8%
查全率	77.59%	68.9%	76.1%
综合值	74.93%	66%	73.88%

第二组对比实验中，本文方法的各个性能指标均优于文献[12]和文献[4]的实验结果。因此本文在保证一定查全率的情况下仍得到了较好的查准率，再次表明了本文方法在特征提取领域的有效性。

5　结　语

本文基于方差选择和 TF-IDF 算法对产品特征进行预抽取；制定名词非特征规则对候选特征进行进一步过滤；采用基于矩阵和权重的改进 Apriori 算法对产品特征进行最优特征挖掘。实验结果表明，与其他特征提取方法相比较，在人工标注的特征较多的情况下，本文方法仍能保持较高的准确率和查全率，说明本文方法是有效的。有效的产品特征是用户做出购买决策的有效参数，也是生产商和销售商改进商品和服务的关键指标，更是在许多商业活动中对产品推荐起到了理想的作用。今后也将结合更多机器学习算法对评论文本中的情感倾向性进行相关研究。

参考文献：

[1] Zhuang L, Jing F, Zhu X Y. Movie Review Mining and Summarization [C]//Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, Virginia, USA.New York: ACM, 2006: 43-50.

[2] Kobayashi N, Inui K, Matsumoto Y, et al. Collecting Evaluative Expressions for Opinion Extraction [C]// Proceedings of the 1st International Joint Conference on Natural Language Processing. Berlin, Heidelberg: Springer-Verlag, 2004: 596-605.

[3] 娄德成，姚天昉. 汉语句子语义极性分析和观点抽取方法的研究[J]. 计算机应用，2006，26(11) : 2622-2625. (Lou Decheng, Yao Tianfang. Semantic Polarity Analysis and Opinion Mining on Chinese Review Sentences[J]. Journal of Computer Applications, 2006, 26(11): 2622-2625.)

[4] Hu M, Liu B. Mining Opinion Features in Customer Reviews [C]// Proceedings of the 19th National Conference on Artificial Intelligence. 2004.

[5] Popescu A M, Etzioni O. Extracting Product Features and Opinions From Reviews [C]//Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005.

[6] 杜思奇，李红莲，吕学强. 汉语组块分析在产品特征提取中的应用研究[J]. 现代图书情报技术，2015(9): 26-30. (Du Siqi, Li Honglian, Lv Xueqiang. Application of Chinese Chunk Analysis in Product Feature Extraction [J]. New Technology of Library and Information Service, 2015(9): 26-30.)

[7] 王永，张勤，杨晓洁. 中文网络评论中产品特征提取方法研究[J]. 现代图书情报技术，2013(12): 70-73. (Wang Yong, Zhang Qin, Yang Xiaojie. Study on the Extraction of Product Features in Chinese Network Reviews [J]. New Technology of Library and Information Service, 2013(12): 70-73.)

[8] 路永和，梁明辉. 遗传算法在改进文本特征提取方法中的应用[J]. 现代图书情报技术，2014(4): 48-57. (Lu Yonghe,

Liang Minghui. Application of Genetic Algorithms in Improving Text Feature Extraction Method [J]. New Technology of Library and Information Service, 2014 (4): 48-57.)
[9] 张建娥. 基于 TFIDF 和词语关联度的中文关键词提取方法[J]. 情报科学, 2012, 30(10): 1542-1544, 1555. (Zhang Jian'e. Chinese Keyword Extraction Method Based on TFIDF and Word Relevance Degree [J]. Information Science, 2012, 30 (10): 1542-1544, 1555.)
[10] 边根庆, 王月. 一种基于矩阵和权重改进的 Apriori 算法[J]. 微电子学与计算机, 2017, 34(1): 136-140. (Bian Genqing, Wang Yue.A Apriori Algorithm Based on Matrix and Weight Improvement [J]. Microelectronics and Computer, 2017, 34 (1): 136-140.)
[11] Shi B, Chang K. Mining Chinese Reviews[C]//Proceedings of the 6th IEEE lnrternational Conference on Data Mining. 2006.
[12] 李实, 叶强, 李一军, 等. 中文网络客户评论的产品特征挖掘方法研究[J]. 管理科学学报, 2009, 12(2): 142-152. (Li Shi, Ye Qiang, Li Yijun, et al. Research on Product Feature Mining Method of Chinese Network Customer Review [J]. Chinese Journal of Management Science, 2009, 12 (2): 142-152.)
[13] 李实, 叶强, 李一军, 等. 挖掘中文网络客户评论的产品特征及情感倾向[J]. 计算机应用研究, 2010, 27(8): 3016-3019. (Li Shi, Ye Qiang, Li Yijun, et al. Characteristics and Emotional Tendency of Excavating Chinese Network Customer Reviews [J]. Application Research of Computers, 2010, 27 (8): 3016-3019.)

作者贡献声明:

李昌兵: 提出研究方向及研究思路, 修改论文;
庞崇鹏: 论文撰写, 实验研究及结果对比分析;
李美平: 英文摘要撰写。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: Pang_Aaron@163.com。
[1] 李昌兵, 庞崇鹏. 1 原始数据集.txt. 将评论语料合并后的数据.
[2] 李昌兵, 庞崇鹏. 2 分词与词性标注.txt. 利用 python 的 jieba 模块进行分词结果.
[3] 李昌兵, 庞崇鹏. 3 去重后特征.txt. 将初步提取出来的特征进行去重.
[4] 李昌兵, 庞崇鹏. 4 取交结果.txt. 利用方差分析与 TFIDF 方法进行结果取交.
[5] 李昌兵, 庞崇鹏. 5 删除非产品特征规则词结果.txt. 在第四步中删除非产品特征规则词.
[6] 李昌兵, 庞崇鹏. 6 改进 Apriori 算法提取最优结果.txt. 将第五步的特征利用 Apriori 算法进行最后过滤得到的结果.

收稿日期: 2017-04-24
收修改稿日期: 2017-06-20

Extracting Product Features with Weight-based Apriori Algorithm

Li Changbing　Pang Chongpeng　Li Meiping
(School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: [**Objective**] This paper aims to reduce the noises while extracting product features from customer comments. [**Methods**] We used the TF-IDF and variance selection methods to extracted the needed data. Then, we set the thresholds to filter the extracted words and obtain the product feature set. Third, we generated frequent item sets with the Apriori algorithm. Finally, we defined various thresholds to obtain the optimal sets, which automatically extracted product features from user comments. [**Results**] We examined the effectiveness of the proposed method with comment texts on mobile phone products. Comparing the automatically extracted characteristics with the manually identified characteristics, we found that the precision P value was 72.44%, the recall R value was 77.59%, and the comprehensive F value reached 74.93%. [**Limitations**] The precision needs to be improved and there might be some human errors involving the manually identified terms. [**Conclusions**] The Apriori algorithm could help us extract product features effectively.

Keywords: Feature Extraction　Apriori Algorithm　TF-IDF　Variance Selection